GGGP-Based Method for Modeling Time Series: Operator Selection, Parameter Optimization and Expert Evaluation

Fernando Alonso
Dept. LSIIS. Fac.Informática, UPM
Boadilla del Monte, Madrid, Spain
+34 91 352 25 46
falonso@fi.upm.es

Loïc Martínez
Dept. LSIIS. Fac.Informática, UPM
Boadilla del Monte, Madrid, Spain
+34 91 352 25 46
loic@fi.upm.es

Agustín Santamaría
Dept. LSIIS. Fac.Informática, UPM
Boadilla del Monte, Madrid, Spain
+34 686 60 58 28
agustin.santamaria@gmail.com

Aurora Pérez
Dept. LSIIS. Fac.Informática, UPM
Boadilla del Monte, Madrid, Spain
+34 91 336 73 89
aurora@fi.upm.es

Juan P. Valente
Dept. LSIIS. Fac.Informática, UPM
Boadilla del Monte, Madrid, Spain
+34 91 336 73 89
jpvalente@fi.upm.es

ABSTRACT
This paper describes the theoretical and experimental analysis conducted to define the best values for the various operators and parameters of a grammar-guided genetic programming process for creating isokinetic reference models for top competition athletes. Isokinetics is a medical domain that studies the strength exerted by the patient joints (knee, ankle, etc.). We also present an evaluation of the resulting reference models comparing our results with the reference models output using other methods.

Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data mining.

General Terms
Experimentation

Keywords
Reference Model, Experimentation, Time Series, Grammar-Guided Genetic Programming, Operators Selection.

1. INTRODUCTION
In a previous article [1], we described a method, called SYRMO, to create models from symbolic time series using GGGP. Taking into account that the symbolic series comply with a number of constraints and that these constraints can be formalized using a Context-Free Grammar (CFG), we thought of using GGGP as a method to create reference models.

A brief description of this process follows. The initial population is composed not of a random population but of the set of isokinetics exercises for which we want to create the reference model. As the process advances, the individuals are selected, crossed and mutated generation by generation to improve the fitness function, taking into account the restrictions imposed by the CFG. The fitness function evaluates each individual's quality as the arithmetic mean of each individual's similarity to the other individuals in the initial population.

In this paper we present the experiments run to define the configuration of the genetic process. This was done through the following phases: a) we conducted a thorough analysis of each of the different operators that can be used in each of the stages of the process; b) as a result of this analysis and several experiments run, we chose the operators that matched the problem to be addressed and rejected others that were of an unsuitable type; c) we ran a range of tests on the set of selected operators to establish the parameterization that returns the best results (the parameters leading to the best operator behavior); d) after choosing the operators and the parameters to be used, we applied the method in different experiments designed to assess the quality of the resulting reference models.

2. GENETIC OPERATORS AND CONVERGENCE CRITERIA
The considered selection operators include the tournament operator, the roulette operator, the scaling operator and the generational operator. None of the above has any theoretical feature leading us to either reject or prefer it over the others. Therefore, we had to run experiments to choose the best selection operator for our problem (see Section 3).

Both the Whigham (WX) [2] and GBX (Grammar-Based Crossover) [3] crossover operators guarantee the generation of valid individuals because they use the CFG underlying the GGGP algorithm. So we used both in the experiments. For the mutation, we used the GBM (grammar-based mutation) [3] operator as it overcomes some restrictions of the standard mutation operator. As replacement operator we choose the SSGA operator.

We weighed up different approaches to convergence. They included the percentage approach (the algorithm converges when a specified percentage of individuals are above the fitness

GECCO'10, July 7–11, 2010, Portland, Oregon, USA.
ACM 978-1-4503-0072-8/10/07.

threshold), the maximum number of generations approach or a mixed convergence method that places time constraints on percentage convergence. As the aim of our tests was to achieve good results without any previous knowledge or restrictions on algorithm execution time, we chose the percentage approach.

3. PARAMETER OPTIMIZATION EXPERIMENTS

We created 10 experimental reference models with three initial population sizes of 10, 20 and 30 individuals merely to adjust the parameters used in the GGGP process and determine which the best genetic operators were in terms of convergence speed and fitness of the generated model. Table 1 summarizes the tests run to adjust the SYRMO parameters[1]. Table 2 lists the number of generations and best fitness for each test.

Table 1. Configuration of each test

Test	Selection	Crossover	Mutation		Replac.	Depth Control	Convergence			
1	Tournament	WX	No Mutation		SSGA	No Control	%	92	FT	90
2	Tournament	WX	No Mutation		SSGA	No Control	%	90	FT	89
3	Tournament	WX	No Mutation		SSGA	Strict L. = 10	%	93	FT	87
4	Tournament	WX	No Mutation		SSGA	Dynamic L. = 10	%	93	FT	87
5	Generational	WX	No Mutation		SSGA	Strict L. = 10	%	95	FT	90
6	Scaling	WX	No Mutation		SSGA	Strict L. = 10	%	95	FT	90
7	Roulette	WX	No Mutation		SSGA	Strict L. = 10	%	95	FT	90
8	Generational	WX	GBM (L=10) Node (Single)	% 10	SSGA	Strict L. = 10	%	98	FT	92
9	Generational	WX	GBM (L=10) Node (Multi)	% 10	SSGA	Strict L. = 10	%	98	FT	92
10	Generational	GBX (L=10)	GBM (L=10) Node (Single)	% 10	SSGA	Strict L. = 10	%	98	FT	92

Table 2. Results for each test

	10 exercises		20 exercises		30 exercises	
Test	Generations	Fitness	Generations	Fitness	Generations	Fitness
1	84	92,130	310	92.656	1001	92.235
2	28	90.186	26	90.184	260	89.954
3	31	93.554	72	92.536	319	93.153
4	263	94.275	96	94.091	335	93.981
5	106	90.272	106	90.202	194	90.042
6	94	90.269	95	90.184	185	90.807
7	117	90.026	130	90.92	223	90.976
8	117	92.784	106	92.719	206	92.923
9	386	92.912	538	92.903	1257	92.97
10	109	92.916	109	92.984	238	92.993

These experiments lead to the conclusion that the best configuration for the reference models creation in SYRMO is as follows: the percentage convergence method, a strict depth limit, the scaling selection operator, GBM with single node mutation as mutation operator and GBX operator as crossover operator.

4. EXPERT EVALUATION OF THE REFERENCE MODELS

To evaluate how good the GGGP-generated symbolic reference models were, we focused on two points: a) check whether the physiotherapist achieved more effective results by analyzing symbolic isokinetics reference models than using numerical isokinetics reference models (built with a numerical method based on the Fourier transform that we used previously); b) check whether the system returned more significant results comparing symbolic sequences using the symbolic isokinetics reference models to classify athletes than comparing their respective numerical sequences using the numerical isokinetics reference models.

We used 20 reference models that had been generated using the symbolic method. These reference models were created from populations of football, basketball and handball players, and swimmers and skaters.

As summary of the experimentation we found that symbolic reference models are slightly more discriminative than numerical reference models because they focus more on the isokinetics concepts that are relevant for the expert.

5. ACKNOWLEDGMENTS

This work was partially funded by the Spanish Ministry of Science and Innovation (MICINN) through the VIIP Project (DEP2005-00232-C03).

6. REFERENCES

[1] Alonso, F., Martínez, L., Pérez, A., Santamaría, A. and Valente, J.P. Symbol Extraction Method and Symbolic Distance for Analysing Medical Time Series, LNCS, 4345, 311-322, 2006.

[2] Whigham, P.A., "Grammatically-Based Genetic Programming", *In Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications*, University of Rochester, Rochester, NY, 33-41, 1995.

[3] Couchet, J., Manrique, D., Ríos, J., and Rodríguez-Patón, A. Crossover and mutation operators for grammar-guided genetic programming. Soft Computing 11, 943-955, 2007.

[1] L parameter is the depth limit used for GBX and GBM.